Efficient Bayesian data assimilation via inverse regression

Benoit KUGLER [1] & Florence FORBES [1] & Sylvain DOUTE [2] & Michel GAY [3]

[1] *Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LJK, 38000 Grenoble, France (firstname.lastname@inria.fr)*
[2] *Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France (sylvain.doute@univ-grenoble-alpes.fr)*
[3] *Grenoble Image sPeech Signal and Automatics Lab, Grenoble 38000 Grenoble, France (michel.gay@gipsa-lab.grenoble-inp.fr)*

Résumé. On propose une approche bayésienne pour résoudre un problème d'assimilation de données. Dans un premier temps, le modèle direct est approché par un modèle paramétrique inversible. Dans un deuxième temps, l'information a-priori est intégrée. Cette division en deux étapes permet de traiter efficacement un nombre important d'inversions. La méthode est illustrée sur une étude du manteau neigeux, utilisant un modèle de rétro-diffusion électro-magnétique.

Mots-clés. Assimilation de données, problème inverse, régression, apprentissage statistique

Abstract.

We propose a Bayesian approach to data assimilation problems, involving two steps. We first approximate the forward physical model with a parametric invertible model, and we then use its properties to leverage the availaibility of a priori information. This approach is particularly suitable when a large number of inversions has to be performed. We illustrate the proposed methodology on a multilayer snowpack model.

Keywords. Data assimilation, inverse problem, regression, statistical learning

1 Introduction

A data assimilation task aims at retrieving unknown parameters, denoted by $\mathbf{x}$, from observations $\mathbf{y}$ and an initial guess on the parameters $\mathbf{x}_0$. The observations and the parameters are linked by a forward model, denoted by F. This problem is similar to inverse problems but differs in the sense that the number of observations $\mathbf{y}$ is much smaller than the number of parameters so that the observations $\mathbf{y}$ alone are not enough to predict the parameters. It is then crucial to make full use of an initial guess of the parameters $\mathbf{x}_0$ to avoid an ill-posed problem. One way to formalize this problem is to adopt a Bayesian formulation. Our problem is modeled considering two random variables $\mathbf{X} \in \mathbb{R}^L$ and $\mathbf{Y} \in \mathbb{R}^D$, linked by the relation

$$Y = F(X) + \varepsilon \tag{1}$$

where ε is a centered Gaussian noise, with variance $\mathbf{\Sigma}$, accounting for the measurement and model uncertainties. We then account for an initial guess $\mathbf{x}_0$ with a prior density on $\mathbf{X}$, for example the product of a Gaussian distribution with mean $\mathbf{x}_0$ and a variance $\mathbf{\Gamma}_0$ and of a uniform distribution on the parameters range, denoted by $\mathcal{U}_{\mathcal{P}}$. According to Bayes' rule, the posterior distribution has then the following form:

$$p^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \propto \mathcal{U}_{\mathcal{P}}(\mathbf{x})\mathcal{N}(\mathbf{x}; \mathbf{x}_0, \mathbf{\Gamma}_0)\,\mathcal{N}(F(\mathbf{x}); \mathbf{y}, \mathbf{\Sigma}) \ .$$

The choice of $\mathbf{\Gamma}_0$ and $\mathbf{\Sigma}$ is crucial. Taking $\mathbf{\Sigma} = 0$ boils down to solve the inverse problem alone, without taking into account prior information. In contrast, taking $\mathbf{\Gamma}_0 = 0$ just yields a dirac centered at $\mathbf{x}_0$, without exploiting the measurements. When looking at the maximum a posteriori (MAP) solution, it comes

$$\hat{\mathbf{x}}_{MAP} = \arg\min_{\mathbf{x} \in P} ||\mathbf{x} - \mathbf{x}_0||_{\mathbf{\Gamma}_0} + ||\mathbf{y} - F(\mathbf{x})||_{\mathbf{\Sigma}} \ ,$$

where $|| \cdot ||_{\mathbf{\Sigma}}$ denotes the Mahalanobis distance. This is a well known consequence of assuming Gaussian distributions for the forward and prior models. In this work, we propose to go beyond the Gaussian assumption by learning the underlying relation between $\mathbf{X}$ and $\mathbf{Y}$, using a regression approach. We first introduce the statistical model and show how it can be used in an assimilation problem. We then illustrate the method on a realistic example in remote sensing.

2 Efficient assimilation via regression

We propose to use a two-steps approach: first, we consider the problem without prior information, and we learn the underlying relation between $\mathbf{X}$ and $\mathbf{Y}$, using the so-called Gaussian Locally-Linear Mapping model (GLLiM) ([Deleforge et al., 2015]). Then, we adapt the model to take into account prior information.

2.1 Learning an invertible approximation of the forward model

In this first step, the joint distribution of $\mathbf{X}$ and $\mathbf{Y}$ is approximated by a Gaussian Locally-Linear Mapping model (GLLiM) which builds upon Gaussian mixture models to capture non linear relationships ([Deleforge et al., 2015]). A latent variable $Z \in \{1, \ldots, K\}$ is introduced to model $\mathbf{Y}$ as piece-wise affine transformation of $\mathbf{X}$:

$$\mathbf{Y} \;\; = \;\; \sum_{k=1}^{K} \mathbb{I}_{\{Z=k\}}(\boldsymbol{A}_k\mathbf{X} + \boldsymbol{b}_k + \boldsymbol{\epsilon}_k) \tag{2}$$

where $\mathbb{I}$ is the indicator function, $\boldsymbol{A}_k$ a $D \times L$ matrix and $\boldsymbol{b}_k$ a vector of $\mathbb{R}^D$ that define an affine transformation. Variable $\boldsymbol{\epsilon}_k$ corresponds to an error term which is assumed to be zero-mean and not correlated with $\mathbf{X}$ capturing both the observation noise and the reconstruction error due to the affine approximation.

In order to keep the posterior tractable, we assume that $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$ and $\mathbf{X}$ is a mixture of K Gaussians : $p(\mathbf{x}|Z = k) = \mathcal{N}(\mathbf{x}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ and $p(Z = k) = \pi_k$. The GLLiM model is thus characterized by the parameters $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1:K}$

This model can be learned from a training set using an EM algorithm. More specifically, the training set $(\mathbf{x}_n, \mathbf{y}_n)_{n=1..N}$ is simulated such that $\mathbf{x}_n$ are realizations of the prior $\mathcal{U}_{\mathcal{P}}(\mathbf{x})$ and $\mathbf{y}_n = F(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$. We then use the resulting GLLiM distribution denoted by p_G (and depending on $\boldsymbol{\theta}$) as a surrogate model for the pdf of $(\mathbf{X}, \mathbf{Y})$. Let's stress out that this first step does not use any prior information on $\mathbf{X}$.

The purpose is to exploit the tractable density p_G provided by the GLLiM model. Indeed, from p_G, conditional distributions are available in closed form and in particular:

$$
\begin{aligned}
p_G(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) &= \sum_{k=1}^{K} w_k^*(\mathbf{y}) \mathcal{N}(\mathbf{x}; \boldsymbol{A}_k^* \mathbf{y} + \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*) \\
\text{with } w_k^*(\mathbf{y}) &= \frac{\pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^{K} \pi_j^* \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j^*, \boldsymbol{\Gamma}_j^*)}
\end{aligned}
\tag{3}
$$

where a new parametrization $\boldsymbol{\theta}^* = \{\boldsymbol{c}_k^*, \boldsymbol{\Gamma}_k^*, \boldsymbol{A}_k^*, \boldsymbol{b}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1:K}$ is used that can be easily deduced from $\boldsymbol{\theta}$ as follows:

$$
\begin{aligned}
\boldsymbol{c}_k^* &= \boldsymbol{A}_k \boldsymbol{c}_k + \boldsymbol{b}_k \\
\boldsymbol{\Gamma}_k^* &= \boldsymbol{\Sigma}_k + \boldsymbol{A}_k \boldsymbol{\Gamma}_k \boldsymbol{A}_k^\top \\
\boldsymbol{\Sigma}_k^* &= \left(\boldsymbol{\Gamma}_k^{-1} + \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{A}_k \right)^{-1} \\
\boldsymbol{A}_k^* &= \boldsymbol{\Sigma}_k^* \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \\
\boldsymbol{b}_k^* &= \boldsymbol{\Sigma}_k^* \left(\boldsymbol{\Gamma}_k^{-1} \boldsymbol{c}_k - \boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1} \boldsymbol{b}_k \right)
\end{aligned}
\tag{4}
$$

The next section shows how to integrate the given prior information on $\mathbf{X}$.

2.2 Prediction step using prior information

We now observe that the target posterior (with prior information) can be factored into the product of the prior and a prior-less posterior:

$$
p^0(\mathbf{x}|\mathbf{Y} = \mathbf{y}, \boldsymbol{\theta}) \propto p(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \mathcal{N}\left(\mathbf{x}; \mathbf{x}_0, \boldsymbol{\Gamma}_0\right)
$$

where

$$
p(\mathbf{x}|\mathbf{Y} = \mathbf{y}) \propto \mathcal{U}_{\mathcal{P}}(\mathbf{x}) \mathcal{N}\left(F(\mathbf{x}); \mathbf{y}, \boldsymbol{\Sigma}\right)
$$

Since the GLLiM model has been learned such as to provide an approximation of $p(\mathbf{x}|\mathbf{Y}=\mathbf{y})$ through (3), we approximate the target posterior with

$$p_G^0(\mathbf{x}|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}) \propto \mathcal{N}(\mathbf{x};\mathbf{x}_0,\boldsymbol{\Gamma}_0)\, p_G(\mathbf{x}|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta}) \tag{5}$$

The key feature is that this density remains in closed form: it actually remains a Gaussian Mixture, with weights, means and covariances $(\alpha_k, \mathbf{x}_k, S_k)_{K=1\ldots K}$ given by

$$\begin{aligned}
S_k &= \left(\boldsymbol{\Gamma}_0^{-1} + (\boldsymbol{\Sigma}_k^*)^{-1}\right)^{-1} \\
\mathbf{x}_k &= S_k\left(\boldsymbol{\Gamma}_0^{-1}\mathbf{x}_0 + (\boldsymbol{\Sigma}_k^*)^{-1} m_k^*\right) \\
\beta_k &= \sqrt{\frac{|S_k|}{(2\pi)^L|\boldsymbol{\Gamma}_0||\boldsymbol{\Sigma}_k^*|}} \exp\left(-\frac{1}{2}(m_k^* - \mathbf{x}_0)^\top B_k (m_k^* - \mathbf{x}_0)\right) \\
B_k &= (\boldsymbol{\Gamma}_0 + \boldsymbol{\Sigma}_k^*)^{-1} \\
\alpha_k &= \frac{w_k^*(\mathbf{y})\beta_k}{\sum_{k=1}^K w_k^*(\mathbf{y})\beta_k} \\
m_k^* &= \boldsymbol{A}_k^*\mathbf{y} + \boldsymbol{b}_k^*
\end{aligned} \tag{6}$$

This means that, once the first learning step is done, inference on the posterior can be performed very efficiently: for example one can solve the assimilation problem by computing the mean of $p_G^0(\mathbf{x}|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta})$, which is straightforward. An uncertainty estimation is also available by computing the variance.

Note that the same formulas can be recovered by observing that accounting for an initial guess $\mathbf{x}_0$ amounts to add in the observations $\mathbf{y}$ an additional observation $\mathbf{x}_0$. Then when using a Gaussian prior for $\mathbf{x}_0$, this combines well with the initial GLLiM model to lead to an *augmented* GLLiM model in dimension $L \times (L+D)$, defined by $(\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)$ being left unchanged and $(\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k)$ modified into

$$\tilde{\boldsymbol{A}}_k = \begin{pmatrix} \boldsymbol{A}_k \\ \mathbf{I}_L \end{pmatrix}, \tilde{\boldsymbol{b}}_k = \begin{pmatrix} \boldsymbol{b}_k \\ 0_L \end{pmatrix}, \tilde{\boldsymbol{\Sigma}}_k = \begin{pmatrix} \boldsymbol{\Sigma}_k & \mathbf{0}_{D,L} \\ \mathbf{0}_{L,D} & \boldsymbol{\Gamma}_0 \end{pmatrix}$$

2.3 Extension to a more complex prior

So far, we have only considered a really simple prior distribution on $\mathbf{X}$. However, the result from the previous section can easily be extended to the case of Gaussian mixtures. Indeed, we can replace $\mathcal{N}(\mathbf{x};\mathbf{x}_0,\boldsymbol{\Gamma}_0)$ by a Gaussian mixture with parameters $(a_i, \mu_i, \Gamma_i)_{i=1..I}$, and still obtain $p_G^0(\mathbf{x}|\mathbf{Y}=\mathbf{y},\boldsymbol{\theta})$ as a Gaussian mixture, this time with $K \times I$ components. Its parameters $(\alpha_{k,i}, \mathbf{x}_{k,i}, S_{k,i})$ are given by the following equations, which are a generalization of (6) :

$$
\begin{aligned}
S_{k,i} &= \left(\Gamma_i^{-1} + (\mathbf{\Sigma}_k^*)^{-1}\right)^{-1} \\
\mathbf{x}_{k,i} &= S_{k,i}\left(\Gamma_i^{-1}\mu_i + (\mathbf{\Sigma}_k^*)^{-1} m_k^*\right) \\
\beta_{k,i} &= \sqrt{\frac{|S_{k,i}|}{(2\pi)^L|\Gamma_i||\mathbf{\Sigma}_k^*|}} \exp\left(-\frac{1}{2}(m_k^* - \mu_i)^\top B_{k,i}(m_k^* - \mu_i)\right) \\
B_{k,i} &= (\Gamma_i + \mathbf{\Sigma}_k^*)^{-1} \\
\alpha_{k,i} &= \frac{w_k^*(\mathbf{y})a_i\beta_{k,i}}{\sum_{k=1}^K \sum_{i=1}^I w_k^*(\mathbf{y})a_i\beta_{k,i}} \\
m_k^* &= \boldsymbol{A}_k^*\mathbf{y} + \boldsymbol{b}_k^*
\end{aligned}
\tag{7}
$$

This flexibility opens the door to more advanced inference tasks. However, in the following, we focus on the simpler case of a Gaussian prior, which is sufficient in the real world scenario we present in the next section.

3 Illustration on a detailed snowpack model

We present an application of our method to an example coming from [Gay et al., 2015], which study the snowpack composition through an electromagnetic backscattering model (EBM). More specifically, initial parameters values are coming from measurements performed manually by experts. The goal is then to refine these initial measurements using information available in the reflectivity measured by a radar. The quantities at stake relate to the composition of the snow layers, namely the snow diameter d_i and its density ρ_i for each layer $i = 1 : L$. Thus, given L layers, the parameters of interest $\mathbf{x}$ are of length $2 * L$. The backscattering measurement y is a scalar related to the parameters through $y = F(\mathbf{x}) = F_{EBM}(d_1, \ldots, d_L, \rho_1, \ldots, \rho_L)$. We refer to [Phan et al., 2014] for more details and the explicit expression of F_{EBM}.

Figure 1 shows the assimilation results for 4 snow carrots. The measurements come the from NoSREx report (measured at X-band, VV polarized, with an incidence angle of 40°). These results are only preliminary, but exhibit two properties that are consistent with previous findings. First, the same pattern for the diameters as in [Gay et al., 2015] is observed: the initial, expert measurements are consistently two high. Second, the density profiles, after assimilation, are increasing with the depth, which is physically sound.

4 Conclusion

We have proposed a Bayesian inversion approach to solve assimilation tasks. We have shown that the inverse regression approach GLLiM could be also adapted to account for a priori knowledge. This framework is especially interesting when we deal when the forward model is fixed, and assimilation is needed for a high number of observations, initial guesses

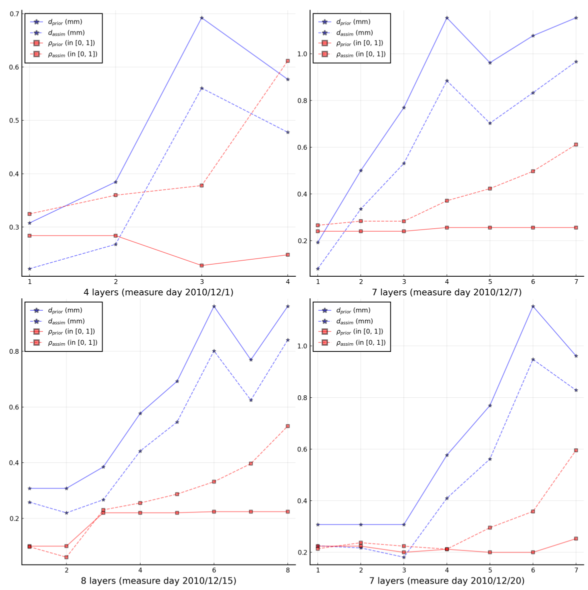

Figure 1: Snow layers properties assimilation. Layers depth is increasing from left to right (that is, the surface is on the left). Snow flakes diameter is in blue, density in red. Initial guesses are in solid line, assimilation result in dashed line.

or prior covariance levels, since the same learned GLLiM model can then be reused. In addition the possibility to use Gaussian mixtures as prior may cover a large range of physical constraints. Future work also includes the study of the covariance choice impact on the final assimilation results.

References

[Deleforge et al., 2015] Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911.

[Gay et al., 2015] Gay, M., Phan, X.-V., Ferro-Famil, L., Karbou, F., Durand, Y., Girard, A., and D 'urso, G. (2015). Simulation de la rétrodiffusion radar du manteau neigeux. Comparaison avec les données d'un radar SOL et TSX (projet NoSREx). In *L'Environnement ElectroMagnétique Des Radars à l'horizon 2020 : Quels Enjeux En Termes de Modélisation et Moyens de Mesures ?*, ENVIREM_FINAL, Gif sur Yvette, France.

[Phan et al., 2014] Phan, X. V., Ferro-Famil, L., Gay, M., Durand, Y., Dumont, M., Morin, S., Allain, S., D'Urso, G., and Girard, A. (2014). 1D-Var multilayer assimilation of X-band SAR data into a detailed snowpack model. *The Cryosphere*, 8(5):1975–1987.